

# De-ID® Update February 2003

## **Project Team**

### ❖ Development by the Center for Biomedical Informatics

- Paul Hanbury
- Melissa Saul
- Greg Cooper, M.D., Ph.D.
- Bruce Buchanan, Ph.D.
- Mehmet Kayaalp M.D., M.S.
- Wendy Chapman, Ph.D.

### ❖ Implementation by the Clinical Research Informatics Service (CRIS)

# De-ID® Mechanics

- De-ID uses a set of heuristics to identify the presence of any of the 18 HIPAA identifiers within the text.
- Supplemental dictionaries of geographic locations, hospital names, and popular names found in the U.S. Census are also used to locate identifiable text.
- The UMLS Metathesaurus is utilized to ensure that words or phrases that may be medical terms containing proper names are preserved.
- De-ID replaces the identifiable text with specific tags. Names found multiple times in the report are consistently replaced with the same tag to improve readability of the report .

# De-ID® Software Development

- Version 1.0 – basic scrubbing of names
- Version 2.0 (May 2001) – handling of for HIPAA defined identifiers as related to text for six commonly used MARS documents
- Version 3.0 (October 2001) – improvements based on CBMI audit
- Version 3.3 (January 2002) – additions to accommodate non-sentence structured documents such as pathology
- Version 4.0 (August 2002) – improvements based on 9 month use; assignment of unique patient ID
- Version 4.14 (January 2003) – refinements in handling of common words and phrases

# The Use of De-ID®

## **Clinical Research within the Schools of the Health Sciences**

- Epidemiology, GSPH
- Health Info Mgmt, SHRS
- Biomedical Informatics, SOM
- Critical Care Medicine, SOM
- ENT, SOM
- Medicine, SOM
- Neurology, SOM
- Orthopedic Surgery, SOM
- Pathology, SOM
- Radiology, SOM
- Outcomes Center, School of Pharmacy
- Acute and Tertiary Care, School of Nursing
- Center for Chronic Diseases, School of Nursing
- Shared Pathology Informatics Network (SPIN) – NCI Funded Collaboration

# The Use of De-ID

## **UPMC Projects**

- UPMC/RAND Women's Health Initiative
- Electronic Health Record Project

## **Other Opportunities**

- Boston area biotechnology company involved in tissue-typing analysis
- VA Healthcare system coordinating a multi-center study
- Healthcare vendor utilizing existing data for benchmarking purposes

# De-ID Usage Statistics

	<b>FY02</b>	<b>FY03YTD</b>
• Datasets de-identified	124	85
• IRB Exempt Projects	51	45
• Formal Audits	3	1

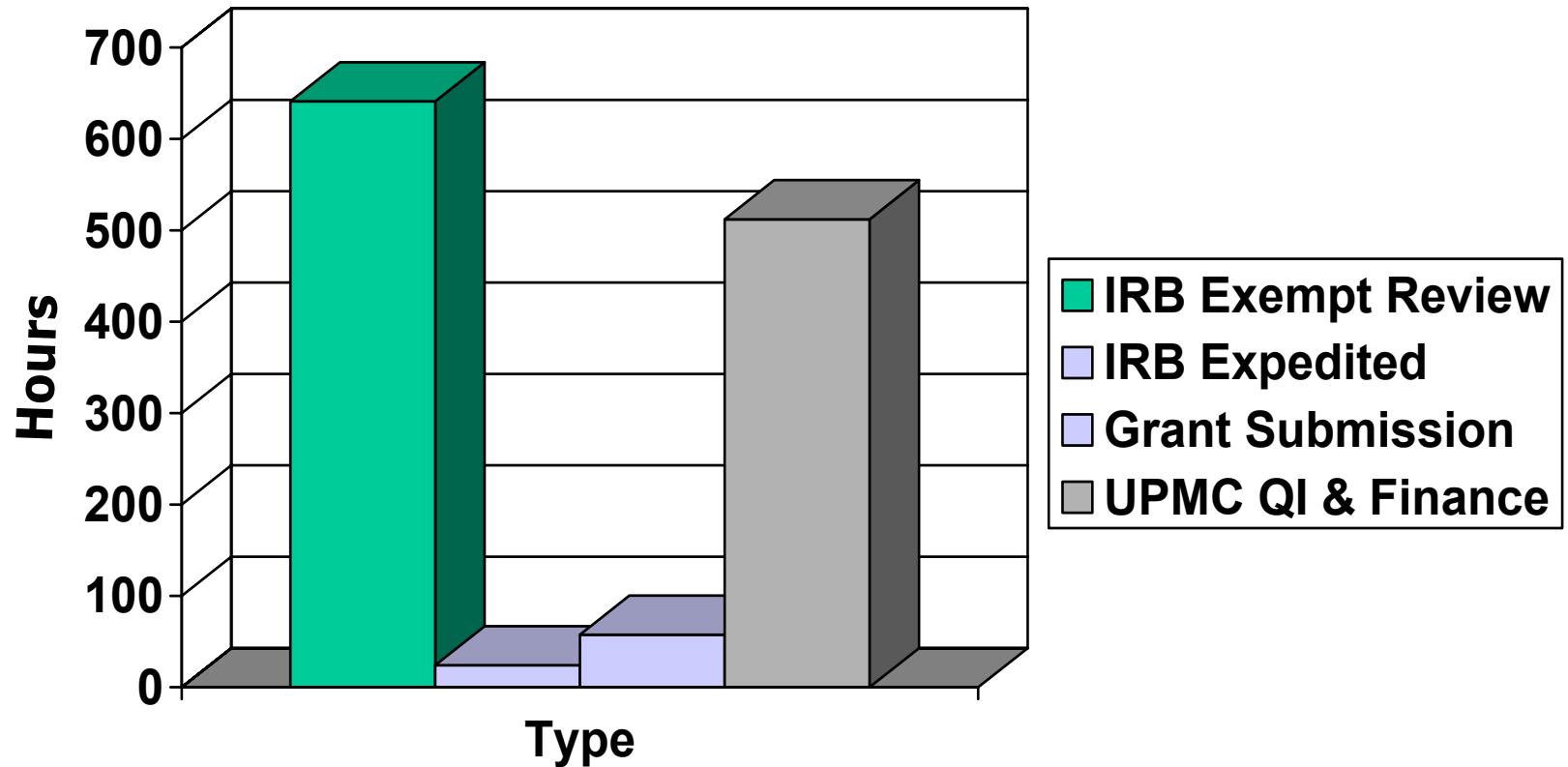
# De-ID in IRB Exempt Studies

## **IRB Application Form, Page 1:**

How will the information be recorded?

*Information electronically de-identified by a computerized system over which research team has no access (e.g., **UPMC (sic) De-ID** program; other program (name):*

# CRIS Projects by Type – FY02





# De-ID® Application

De-ID version 4.14

File Options Help

Original text  
C:\NIPS\mhra\strokeds.out  
Format: MARS-BSH

Output text  
C:\NIPS\mhra\strokeds.dmf  
[Safe-harbor compliant text](#)

Linkage information  
C:\NIPS\mhra\strokeds.dli  
Decryption Pw:

Priority:

Low:  
De-ID's priority is a bit below the default.

Available Header Fields:

CKSM	Checksum
ID	Main Patient Identifier
NA	Patient Name
DAT	Principal Date
PQ	Parsing Queue Transaction Number
DOC	Principal Physician
ADM	Subgroup Classifier
EXAM	Exam Mnemonic
STYP	Record Subtype
SPNO	Accession Number
ACCT	Account Number
CTYP	Case Type (Surgical Pathology)
IDS	Alternate Patient ID Numbers
LOC	Location Code
UNIQ	Unique Record Number

Selected Fields:

TYP	Record Type
-----	-------------

De-ID copyright (c) 1999-2003, University of Pittsburgh. All rights reserved. Compiled: Feb 7 2003

# De-ID Options

De-ID options: The 18 HIPAA identifiers [CFR 164.514(b)(2)(i) and 164.514(c)(2)]

<input checked="" type="checkbox"/> Names [1] <ul style="list-style-type: none"><li><input checked="" type="checkbox"/> Patient, relative, employer or household members [1]</li><li><input checked="" type="checkbox"/> Health care providers [1]</li></ul>	<input checked="" type="checkbox"/> Social security [7], medical record [8], health plan [9] and other [10] account numbers
<input checked="" type="checkbox"/> Geographical subdivisions smaller than a state [2] <ul style="list-style-type: none"><li><input checked="" type="checkbox"/> Street addresses [2]</li><li><input checked="" type="checkbox"/> Cities, states and zip codes [2]</li><li><input checked="" type="checkbox"/> Hospital names [2]</li></ul>	<input type="checkbox"/> License numbers [11] and vehicle identifiers [12]
<input checked="" type="checkbox"/> Dates and ages <ul style="list-style-type: none"><li><input checked="" type="checkbox"/> All elements of date, except year [3]</li><li><input checked="" type="checkbox"/> Ages under 90 years [not req'd]</li><li><input checked="" type="checkbox"/> Ages 90 years and over [3]</li></ul>	<input checked="" type="checkbox"/> Device identifiers and serial numbers [13]
<input checked="" type="checkbox"/> Telephone [4] and fax [5] numbers	<input checked="" type="checkbox"/> Web universal resource locators (URLs) [14] and internet protocol (IP) addresses [15]
<input checked="" type="checkbox"/> Electronic mail addresses [6]	<input type="checkbox"/> Biometric identifiers [16] and full-face photographs [17]
	<input type="checkbox"/> Any other uniquely identifying number, characteristic or code [18]

Safe-harbor defaults      Limited dataset defaults

Output format:       OK

# DE-ID Limited Dataset Defaults

De-ID options: The 18 HIPAA identifiers [CFR 164.514(b)(2)(i) and 164.514(c)(2)]

<input checked="" type="checkbox"/> Names [1] <ul style="list-style-type: none"><li><input checked="" type="checkbox"/> Patient, relative, employer or household members [1]</li><li><input type="checkbox"/> Health care providers [1]</li></ul>	<input checked="" type="checkbox"/> Social security [7], medical record [8], health plan [9] and other [10] account numbers
<input checked="" type="checkbox"/> Geographical subdivisions smaller than a state [2] <ul style="list-style-type: none"><li><input checked="" type="checkbox"/> Street addresses [2]</li><li><input type="checkbox"/> Cities, states and zip codes [2]</li><li><input type="checkbox"/> Hospital names [2]</li></ul>	<input type="checkbox"/> License numbers [11] and vehicle identifiers [12]
<input type="checkbox"/> Dates and ages <ul style="list-style-type: none"><li><input type="checkbox"/> All elements of date, except year [3]</li><li><input type="checkbox"/> Ages under 90 years [not req'd]</li><li><input type="checkbox"/> Ages 90 years and over [3]</li></ul>	<input checked="" type="checkbox"/> Device identifiers and serial numbers [13]
<input checked="" type="checkbox"/> Telephone [4] and fax [5] numbers	<input checked="" type="checkbox"/> Web universal resource locators (URLs) [14] and internet protocol (IP) addresses [15]
<input checked="" type="checkbox"/> Electronic mail addresses [6]	<input type="checkbox"/> Biometric identifiers [16] and full-face photographs [17]
	<input type="checkbox"/> Any other uniquely identifying number, characteristic or code [18]

Safe-harbor defaults    Limited dataset defaults

Output format: XML    OK

# XML Formatted Output Option

De-ID options: The 18 HIPAA identifiers [CFR 164.514(b)(2)(i) and 164.514(c)(2)]

<input checked="" type="checkbox"/> Names [1]	<input checked="" type="checkbox"/> Social security [7], medical record [8], health plan [9] and other [10] account numbers
<input checked="" type="checkbox"/> Patient, relative, employer or household members [1]	<input type="checkbox"/> License numbers [11] and vehicle identifiers [12]
<input checked="" type="checkbox"/> Health care providers [1]	<input checked="" type="checkbox"/> Device identifiers and serial numbers [13]
<input checked="" type="checkbox"/> Geographical subdivisions smaller than a state [2]	<input checked="" type="checkbox"/> Web universal resource locators (URLs) [14] and internet protocol (IP) addresses [15]
<input checked="" type="checkbox"/> Street addresses [2]	<input type="checkbox"/> Biometric identifiers [16] and full-face photographs [17]
<input checked="" type="checkbox"/> Cities, states and zip codes [2]	<input type="checkbox"/> Any other uniquely identifying number, characteristic or code [18]
<input checked="" type="checkbox"/> Hospital names [2]	
<input checked="" type="checkbox"/> Dates and ages	
<input checked="" type="checkbox"/> All elements of date, except year [3]	
<input checked="" type="checkbox"/> Ages under 90 years [not req'd]	
<input checked="" type="checkbox"/> Ages 90 years and over [3]	
<input checked="" type="checkbox"/> Telephone [4] and fax [5] numbers	
<input checked="" type="checkbox"/> Electronic mail addresses [6]	

Safe-harbor defaults      Limited dataset defaults

Output format: XML OK

# Sample De-ID Report

## REASON FOR ADMISSION:

This is a \*\*AGE[in 60s]-year-old male with a past medical history significant for hypertension, who developed a throat discomfort while swimming. The patient was admitted on \*\*DATE[May 17 2001]. The patient was admitted with crescendo angina and positive troponin 1.

## HOSPITAL COURSE:

The patient was admitted to a monitored bed, cardiology was consulted and the patient was seen. The patient was placed on aspirin and heparin. The patient was also placed on a beta-blocker. Dr. \*\*NAME[VVV UUU] from cardiology saw the patient and recommended a cardiac catheterization. On \*\*DATE[May 18 2001], the patient had a cardiac catheterization,

# Formal Audit Studies

- Initial evaluation study in 2001 of 350 reports from UPMC PUH
- Study of 967 surgical pathology reports from nine UPMC-HS hospitals from 2000
- Study of 1000 surgical pathology reports from UPMC PUH from sign outs of years 1985, 1990, 1995, 2000
- Current study of 750 reports from UPMC PUH evaluated by experienced physicians

# Refinement Process

- Goal: to identify and remove all phone numbers
  - Formats
    - 3-3-4
    - 3-4
    - 3\*7
    - 7 with a .
    - 10 with a .
- Issues
  - ✓ 800-1000 cc of fluid
  - ✓ FAX # \*4125551212